Data Analysis

No need to underline the importance of understanding statistics and data analysis– Many example in recent news proves that it is crucial to be able to understand what does all this mean- Election results, statistics on unemployment, on global warming, ... **Everything** seems to be based on data analysis. A small story of measurement before starting.

John is a carpenter who wants to install a door in a doorway. He wants then to know how high is the doorway. First he could just look at it and estimates that it is 210 cm high. That's fine, but for the door installation he would need a more precise measurement. So OK, he gets some tape measure and gets a 211.3 cm measurement. That is more **precise** than his original eye-balling estimate, but it's also has some uncertainty. At this point there may be some source of errors that could make the measurement better. For example, maybe John used a tape which is only graduated in half-centimeters. So the mark was between 211.0 and 211.5 and not exactly in the middle and John estimated that it was 211.3 cm. This could get corrected by getting a better tape and make the measurement more accurate. Let's say John is a very obsessive carpenter and wants **NO** uncertainty in his measurement. Here he

goes to buy a laser interferometer to measure the high of this door. Now, the precision of the measurement is going to be limited by the wavelength of the light used in the measure (about 5×10^{-7} m), so even then John would not know the height of the door **exactly**. Furthermore, he is going to find that he cannot even define a single quantity as being **the height** of the door: the height will vary at some places the paint is a little more bulkier, as some others there are scratches,.... The height of the door is not a well-defined quantity. This is called a **problem of definition** and comes into play in many scientific measurements.

The morale of the story: It is impossible to eliminate uncertainties completely- understanding and reducing them becomes the goal.

A) Mean, variance, skewness and so forth

We start from a set of values x_i, where i varies from 1 to N– To characterize the distribution of x_i, we can compute few numbers which are related to its different "moments" M_n (sums of integer powers of the values).
The mean is linked to M_1:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}M_1 \tag{1}$$

The mean estimate the value around which the x_i distribution clusters.
The variance is linked to M_2 and M_1:

$$\text{Var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{2}$$

and the standard deviation: $\sigma = \sqrt{Var}$
Note: With this definition, σ is called "sample standard deviation" . When the denominator $N-1$ is changed to N, σ is called "population standard deviation". The denominator $N-1$ should be changed to N if you're measuring the variance of a distribution for which you know **a priori** $\overline{x}$ the mean (ie: you don't estimate it from the data).
The standard deviation estimates the mean deviation of the distribution from the mean value.

The skewness:

$$\text{Skew} = \frac{1}{N}\sum_{i=1}^{N}[\frac{x_i - \overline{x}}{\sigma}]^3 \tag{3}$$

The skewness characterizes the degree of asymmetry of a distribution around its mean.

Higher orders quantities can be defined by a general formula:

$$Z_n = \frac{1}{N}\sum_{i=1}^{N}[\frac{x_i - \overline{x}}{\sigma}]^n \tag{4}$$

B) Accuracy, precision, uncertainties, and so on

Accuracy refers to the closeness of the measurements to the value of a physical quantity, whereas the term precision is used to indicate the closeness with which the measurements agree with one another quite independently of any systematic error involved.

C) Random, statistical and systematic errors

Random (statistical) error: associated with any statistical process-

Systematic error (uncertainty): error that will occur **no matter what** as a result of the instrument used to make the measurement-

Quick summary on propagation of errors

Let δx be the error on the measure of a quantity x.
We have (measured x) == $\mathrm{x_{best}} \pm \delta x$
δx is not telling the complete story: a δx of 1cm on a quantity about 10cm is not the same thing that a δx of 1cm on the height of Everest. This is why we introduce the notion of **fractional uncertainty** which is simply $\frac{\delta x}{|x_{\mathrm{best}}|}$.

A) Errors on a sum or a difference

1) First approximation:
If $S = x_1 + x_2 + x_3 + ... + x_N$, the $\delta_S = \delta_{x_1} + \delta_{x_2} + ... \ \delta_{x_N}$
2) If all the x_i are independent and random:
$\delta_S = \sqrt{\delta_{x_1}^2 + \delta_{x_2}^2 + ... \delta_{x_N}^2}$

B) Errors on a product or a quotient

1) First approximation:
If $P = \frac{x_1 \times x_2 \times ... \times x_N}{y_1 \times y_2 \times ... \times y_M}$
$\delta_P = \frac{\delta_{x_1}}{x_1} + ... + \frac{\delta_{x_N}}{x_N} + \frac{\delta_{y_1}}{y_1} + ... + \frac{\delta_{y_M}}{y_M}$
2) If all the x_i and y_i are independent and random:
$\frac{\delta_P}{P} = \sqrt{\frac{\delta_{x_1}}{x_1}^2 + ... + \frac{\delta_{x_N}}{x_N}^2 + \frac{\delta_{y_1}}{y_1}^2 + ... + \frac{\delta_{y_M}}{y_M}^2}$

C) Errors on a function of several variables

If x_1, x_2,...,x_N are measured with uncertainties δ_{x_1} ,δ_{x_2},...,δ_{x_N}.
If the uncertainties in x_1, x_2,...,x_N are independent and random, then the uncertainties in f(x_1, x_2,...,x_N) is

$$\delta_f = \sqrt{\left(\frac{\partial f}{\partial x_1}\delta_{x_1}\right)^2 + ... + \left(\frac{\partial f}{\partial x_N}\delta_{x_N}\right)^2}$$

Binomial distribution; Testing an hypothesis

Supposed that a manufacturer of ski waxes claims to have developed the best of the best wax- How do you test this claims?

You could organize races between skis with and without wax and see how many of the waxed skis actually win the races.

To test the hypothesis that the wax does not make any difference, one can ask "What is the probability of wining w of the N races?"

P(w wins in N races)= $\frac{N!}{w!(N-w)!}0.5^N$

(P(N wins in N races)= 0.5^N – if one wants this to be "highly significant" (ie with a probability less than 1%, then one needs more than 7 races).

P follows a binomial distribution– general form: P(w wins in N races)= $\frac{N!}{w!(N-w)!}p^w(1-p)^{N-w}$

where p is the probability of "success".

Important things about the binomial distribution:

1) $\overline{w}$=Np
2) σ_w=$\sqrt{Np(1-p)}$
3) When N becomes large, binomial distribution converges toward Gaussian distribution

Poisson distribution

Poisson distribution describes the results of experiments where events occur at random but at a definite average rate (example: counting the number of electrons emitted by radioactive decays in a given period of time).
In the example given above, the number of electrons counted during a given (and fixed) period of time will most certainly vary. This variation is linked to the physical process of the radioactive decays themselves.
Each radioactive nucleus has a certain probability for decay, but they all decay at a random time.
So now the question becomes: " If we repeat the counting experiment N times, what distribution of the number of decays recorded would follow? "
It turns out that the answers is:
P(w counts any fixed time interval)= $e^{-\mu}\frac{\nu^w}{w!}$

Important things about the Poisson distribution:

1) $\overline{w}=\nu$ ie the parameter of the Poisson distribution is precisely the number expected from a large number of experiments.
2) $\sigma_w=\sqrt{\nu}$
3) When ν becomes large, Poisson distribution converges toward Gaussian distribution

Gaussian or Normal distribution

Continuous distribution– (as opposed to the discrete distributions mentioned above) – Describe the distribution of measurements subjects to many sources of error that are all random and small.
$f_{x_o,\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-x_o)^2/2\sigma^2}$

Important things about the Gaussian distribution:

1) It is the most important distribution in statistics
2) $\overline{x}=x_o$
3) $\sigma_x=\sigma==$ the width of the distribution

Rejection of data – Chauvenet's criterion

Suppose we make N measurements of the same quantity x. From all N measurements, we compute $\overline{x}$ and σ_x. If one measurement x_1 is very far from $\overline{x}$, we compute the quantity $t=\frac{(x_1-\overline{x})}{\sigma_x}$, which is the number of standard deviations by which x_1 differs from $\overline{x}$. Then, we compute the probability that the measurement lies outside interval defined by $\overline{x} \pm t\sigma_x$.

Then n is the total number of measurements N multiplied by the probability. If n is less than 0.5, x_1 value failed the Chauvenet's test and should be rejected.
Many scientists think that data should **never** be rejected (unless there is external evidence that the data have been corrupted in some ways). A more moderate interpretation of the Chauvenet's criterion is that it is useful to identify the statistical "loners" .

Some others distributions

- Multinomial: Gives probability of exactly n_i outcomes of event i, for $i = 1, 2, ...,$ n in N independent trials when the probability of event i in a single trial is a constant.
- Hypergeometric: Gives probability of picking exactly w good units in a sample of N units from a population of M units when there are k bad units in the population
- Geometric: Gives probability of requiring exactly x binomial trials before the first success is achieved
- Pascal: Gives probability of exactly x failures preceding the n success — Sometimes identified as "Negative binomial"
- and so on...

χ^2 minimization

We want to decide if the measured distribution is consistant with the model (theoretical) distribution. To do this we use what is called the χ^2 test. We performe a total of N observations
O_k is the observed value of the kth observation.
E_k are the expected value of kth observation.
First definition:

$$\chi^2 = \Sigma_{k=1}^{N} \frac{(O_k - E_k)^2}{E_k}$$

If $\chi^2 \simeq N$ then the distributions are consistent
if $\chi^2 \gg N$ the two distributions are inconsistent.

In fact the **correct** number to which to compare χ^2 is not N (the number of bins/trials) but what is called **degrees of freedom** d. d=N-c where c is the number of number of parameters computed from the data and used in the calculation (c can also be viewed as the number of constraints on the model).

We define the "reduced χ^2" $= \tilde{\chi^2}$=χ^2/d
The measured distribution can be rejected at the CL% Confidence Level, if $\mathrm{P}(\tilde{\chi^2} \geq \tilde{\chi^2_o}) \leq$ CL%
where $\mathrm{P}(x)$ is probability of x and $\tilde{\chi^2_o}$ is the measured reduced χ^2.

The exact formula for $\mathrm{P}(\tilde{\chi^2} \geq \tilde{\chi^2_o})$ is given by

$$P(\tilde{\chi^2} \geq \tilde{\chi^2_o}) = \frac{2}{2^{d/2}\Gamma(d/2)} \int_{\chi^2_o}^{\infty} x^{d-1}\, e^{-x^2/2}\, dx$$

Satellite orbit

1) Target visibility
2) Thermal environment – Reflected sunlight from Earth
3) Electromagnetic environment (radiations, contaminations by scattered solar X-rays
4) Particle environment – cosmics rays, protons, ...
5) Zero gravity
6) Vacuum environment – some ice build up on detectors
7) Clocks – each satellite has its own internal clock and light travel time taken into account

Spectral Imaging

1) IRAF/PROS
2) XSPEC
3) CIAO
4) Anything that works...

Timing

1) IRAF/PROS
2) XRONOS
3) Anything that works...

Homework

1) Consider an election between 2 candidates A and B. Suppose that candidate A claims that extensive research has established that he is favored by 60% of the population. B is not happy to hear that and ask you to check out this claim. You select a randomly selected pool of 600 voters and ask their preferences. 330 people says that they prefer candidate A- Can we claim to have cast a significant doubt on the hypothesis that 60% favors A?

2) Problem appended– (number 12.14)